FEBS 27614 FEBS Letters 552 (2003) 207–213

High-throughput screening of structural proteomics targets using NMR

Leonor M.P. Galvão-Botton[a,1], Ângela M. Katsuyama[a,1], Cristiane R. Guzzo[a], Fábio C.L. Almeida[b], Chuck S. Farah[a,*], Ana Paula Valente[b,*]

[a] *Departamento de Bioquímica, Instituto de Química, Universidade de São Paulo, CEP 05508-900, São Paulo, SP, Brazil*
[b] *Departamento de Bioquímica Médica, ICB, CCS, Universidade Federal do Rio de Janeiro, CEP 21941-590, Rio de Janeiro, RJ, Brazil*

Received 3 July 2003; revised 12 August 2003; accepted 14 August 2003

First published online 28 August 2003

Edited by Robert B. Russell

Abstract We applied a high-throughput strategy for the screening of targets for structural proteomics of *Xanthomonas axonopodis* pv *citri*. This strategy is based on the rapid ^{1}H–^{15}N HSQC NMR analysis of bacterial lysates containing selectively ^{15}N-labelled heterologous proteins. Our analysis permitted us to classify the 19 soluble candidates in terms of 'foldedness', that is, the extent to which they present a well-folded solution structure, as reflected by the quality of their NMR spectra. This classification allowed us to define a priority list to be used as a guide to select protein candidates for further structural studies.

Key words: Structural proteomics; NMR; X-ray; Protein screening; Target ORF; *Xanthomonas citri*

1. Introduction

Structural proteomics is a rapidly developing field in biology [1–4]. It involves the determination of protein structures on a genome-wide scale and is expected to yield invaluable information about protein folding and protein function. Several recent examples have demonstrated the feasibility of obtaining functional information through structure [5–7].

Structural proteomics requires a large-scale approach to structural biology and, although rapidly evolving, still presents important bottlenecks that need to be overcome [8,9]. Methods have been developed for high-throughput cloning, expression and purification of target open reading frames (ORFs) [9–12]. Important developments have also been made in automated procedures for X-ray crystallography [12–14] and nuclear magnetic resonance (NMR) spectroscopy structure determination [15–19]. However, important issues, such as the acquisition of well-diffracting crystals and size limitation in NMR, decrease the speed and efficiency of the process. Furthermore, the study of uncharacterized proteins inevitably begins with a large number of candidates that are progressively eliminated due to problems related to expression levels, solubility and stability.

A major challenge of structural proteomics is to select the target proteins in a realistic, efficient and cost-effective way. In order to optimize this process, it would be desirable to have a rapid and efficient method for screening target ORFs for their suitability for structural studies. Several methods have been developed to screen candidates for structural studies, such as cell free expression systems [20,21] and enzymatic labeling of glutamine in vitro [22] and selective isotopic labeling of the target proteins in vivo [23].

In this paper we describe the use of a high-throughput strategy for screening structural proteomics targets by applying it to a set of 35 previously uncharacterized ORFs of the recently sequenced *Xanthomonas axonopodis* pv *citri* (*X. a.* pv *citri*) phytopathogen [24]. This strategy is based on the NMR analysis of bacterial lysates containing selectively ^{15}N-labelled heterologous proteins as originally described by Almeida et al. [23]. By omitting the purification step, this protocol is a fast, efficient and inexpensive method to select proteins with conformational characteristics amenable to further structural analysis by NMR and/or X-ray crystallography.

2. Materials and methods

2.1. Cloning and expression

Targets were amplified by PCR from *X. a.* pv *citri* genomic DNA [24], subcloned into the pET-3a vector and expressed in *Escherichia coli* strain BL21(DE3)pLysS [25]. Cells were grown at 37°C in unlabelled M9 medium up to an $OD_{600\,nm}$ of 0.8 and heterologous protein expression induced by 1 mM isopropyl-β-D-thiogalactoside (IPTG). After 15 min, rifampicin was added to a concentration of 200 μg ml^{-1}. Upon 15 min of further incubation, the cells were pelleted and resuspended in M9 medium containing 1 g l^{-1} $^{15}NH_4Cl$, 1 mM IPTG and 200 μg ml^{-1} rifampicin. Cells were grown for 3.5 h before harvesting and storage at −70°C.

2.2. NMR screening

Frozen cell pellets were thawed by adding 500 μl of 20 mM sodium phosphate (pH 7.0), 10 mM β-mercaptoethanol, 40 μg ml^{-1} PMSF and lysed by sonication. Lysates were clarified by centrifugation and the soluble fraction was concentrated using a 3 kDa cutoff Centricon filter. 10% D_2O was added to the sample. Standard 1D ^{1}H spectra, 1D ^{15}N-edited ^{1}H spectra and 2D ^{1}H–^{15}N heteronuclear single quantum coherence (HSQC) spectra were collected at 303 K in a Bruker Advance DRX 600 MHz. XACb0070 HSQC data were also acquired on a Varian INOVA 600 MHz instrument at the Laboratório Nacional de Luz Sincrotron in Campinas, Brazil. All NMR experiments took an average of 8 h for each sample.

*Corresponding authors.
E-mail addresses: chsfarah@iq.usp.br (C.S. Farah), valente@cnrmn.bioqmed.ufrj.br (A.P. Valente).

[1] These authors contributed equally to this work.

Abbreviations: ORF, open reading frame; NMR, nuclear magnetic resonance; HSQC, heteronuclear single quantum coherence

0014-5793/03/$22.00 © 2003 Published by Elsevier B.V. on behalf of the Federation of European Biochemical Societies.
doi:10.1016/S0014-5793(03)00926-8

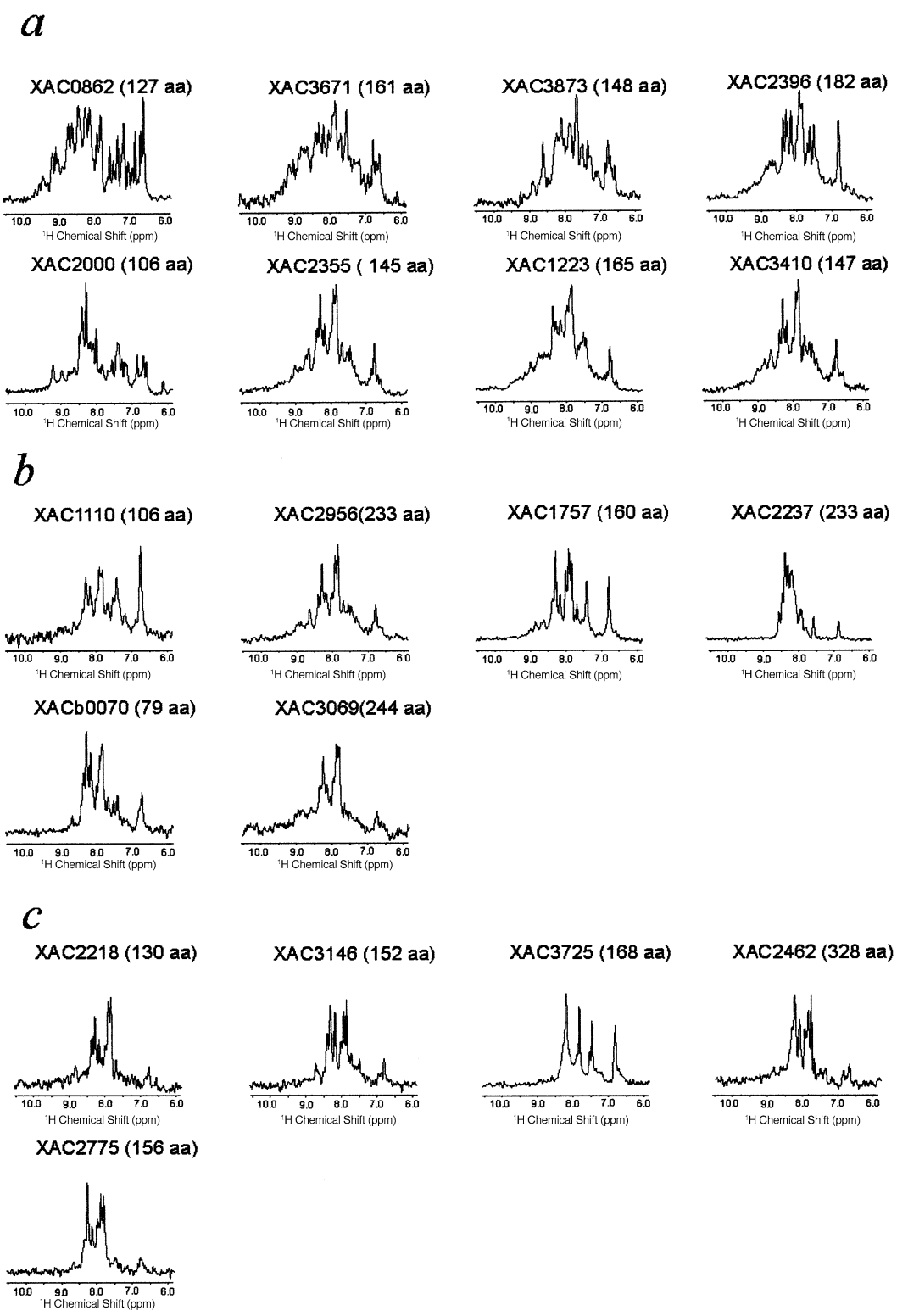

Fig. 1. 1D ^{15}N-edited^{1}H HSQC spectra of the 19 soluble proteins obtained from soluble cell lysates after selective ^{15}N-labelling. Spectra are grouped as *good* (a), *promising* (b) and *poor* (c). Classification is based on the analysis of chemical shift dispersion and line-width. The number above the spectra identifies the ORF encoding the expressed protein with the number of amino acid residues shown in parentheses.

3. Results and discussion

3.1. Selection of target ORFs

Thirty-five *X. a.* pv *citri* gene products [24] were selected on the basis of size (79–330 amino acids), predicted cytosolic localization using PSORT [27], methionine content ($>1.1\%$), low sequence relatedness to proteins of known function and the presence of homologs in the genomes of other organisms (so-called 'conserved hypothetical proteins') [27]. Targets were also selected for absence of homologs in the

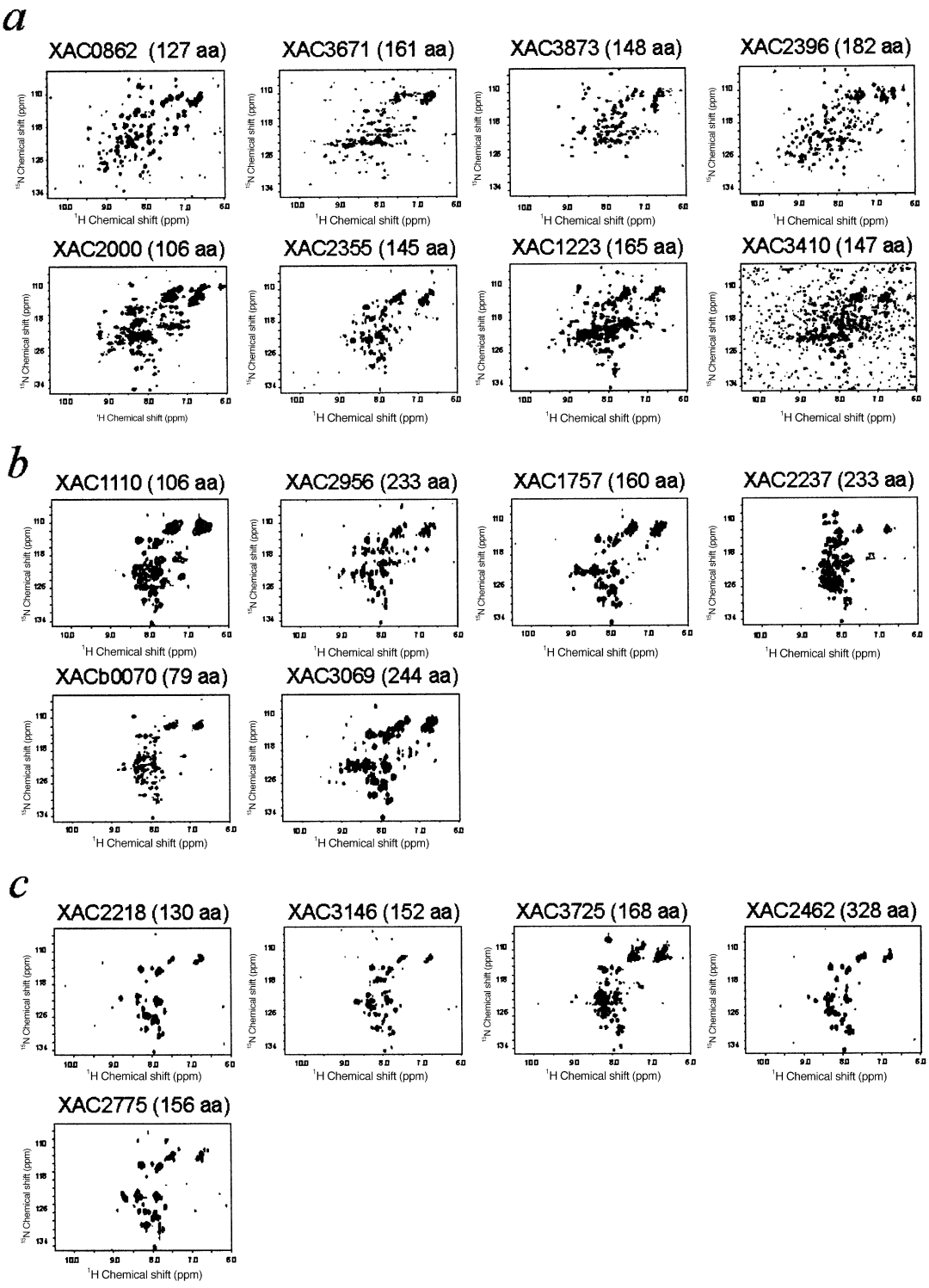

Fig. 2. 2D ^{15}N-edited ^{1}H HSQC spectra of the 19 soluble proteins obtained from soluble cell lysates after selective ^{15}N-labelling. Spectra are grouped as *good* (a), *promising* (b) and *poor* (c). Classification is based on the analysis of chemical shift dispersion, line-width and number of peaks. Spectra are presented in the same order as shown in Fig. 1. The number above the spectra identifies the ORF encoding the expressed protein with the number of amino acid residues shown in parentheses.

PDB (Protein Data Bank, identified using PSI-BLAST [28] according to the method of Huynen et al. [29].

3.2. Expression and solubility

The 35 selected ORFs were amplified and expressed as described in Section 2 [25]. Expression levels and solubility were monitored using SDS–PAGE (sodium dodecyl sulfate–polyacrylamide gel electrophoresis) by the intensity of the induced band present in the cell lysate, pellet and supernatant. Most proteins (31/35) expressed well in *E. coli* grown in M9 minimum medium, and 19 of the expressed proteins remained in the soluble fraction of the cell lysates.

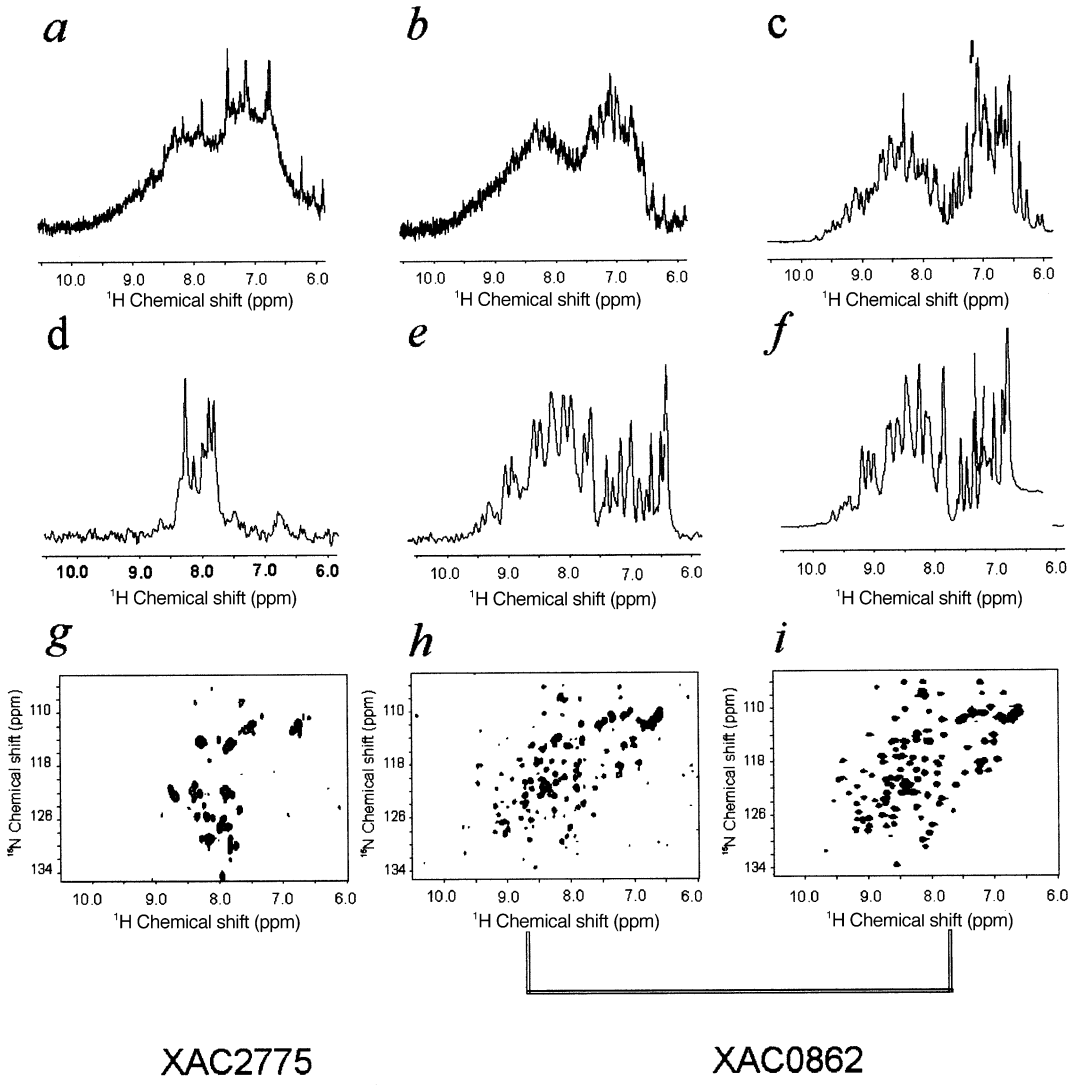

Fig. 3. ^{1}H spectra (a,b,c), ^{15}N-edited ^{1}H spectra (d,e,f) and 2D ^{15}N-edited ^{1}H HSQC spectra (g,h,i) for the proteins encoded by ORFs XAC2775 and XAC0862: soluble bacterial lysates containing ^{15}N-labelled proteins encoded by ORF XAC2775 (a,d,g) and ORF XAC0862 (b,e,h); purified XAC0862protein product (c,f,i).

3.3. Selective ^{15}N-labeling

The 19 soluble proteins were selectively labelled with ^{15}N as described by Almeida et al. [23]. This methodology uses the antibiotic rifampicin to selectively label heterologous proteins for NMR studies. By 'selective' we mean to indicate that only the heterologously expressed protein was labeled with ^{15}N. Rifampicin is a bacterial RNA polymerase inhibitor that does not inhibit T7 polymerase used for expression of heterologous proteins cloned under the control of T7 promoters, as in the case of pET vectors [25]. Thus, heterologous protein expression induction following transfer of the bacteria into isotopically labelled minimal medium supplemented with rifampicin enables the selective labelling of heterologous proteins with an isotope of choice.

3.4. Screening by ^{1}H–^{15}N HSQC

The soluble fractions of the 19 cell lysates containing the selectively ^{15}N-labelled heterologous proteins were rapidly screened by 1D (Fig. 1) and 2D ^{1}H–^{15}N HSQC (Fig. 2). It has long been recognized that chemical shift contains structural information [30,31]. Furthermore, NMR spectra of well-folded proteins exhibit sharp lines and large chemical shift dispersion, while unfolded proteins present spectra with small chemical shift dispersion and often contain broad lines, due to conformational exchange. Thus, the combined analysis of chemical shift dispersion and line-width provides information about the conformational state of the protein in solution, or its 'foldedness', which could be used as criteria for selecting good protein candidates for further structural analysis [31,32,33].

Fig. 1 shows the 1D HSQC spectra obtained for all 19 *X. a.* pv *citri* protein targets screened in this study. According to the quality of the spectra proteins could be empirically grouped. The spectra obtained were classified as *good* when showing sharp, intense and well-dispersed peaks characteristic of well-structured, stable proteins (Fig. 1a). Spectra showing broader, less intense and less dispersed peaks than expected for a stable and well-structured protein were classified as *promising* (Fig. 1b). Such spectra often indicate conformation heterogeneity or dynamic processes on a slow time scale that

can broaden the NMR signal. *Poor* spectra were characterized by a cluster of broad peaks with low chemical shift dispersion (Fig. 1c). These spectra are likely to represent unfolded, conformationally unstable or aggregated proteins.

The 2D HSQC NMR spectra for all 19 soluble proteins are presented in Fig. 2 and again ordered according to the quality of the spectra. The conclusions derived from the 2D HSQC spectra are consistent with those from the 1D HSQC experiments. In addition, in the 2D spectra one can compare the observed number of peaks to that expected according to the primary sequence. As the 2D HSQC spectra are more informative than the 1D spectra, the final priority list as shown in Figs. 1 and 2 was built based mostly on the 2D spectra.

Fig. 3 compares the ^{1}H spectra and the HSQC spectra of the cell lysates after expression of two proteins: XAC2775 (Fig. 3a,d,g) and XAC0862 (Fig. 3b,e,h), which presented *poor* and *good* spectra, respectively. Fig. 3 also shows the spectra of the purified XAC0862 protein (Fig. 3c,f,i). The ^{1}H spectra of the two lysate samples (Fig. 3a,b) are very similar as resonances from all the soluble components in the cell lysate appear. In contrast, the HSQC spectra present fewer and more defined peaks that correspond to the amide protons of the ^{15}N-labelled heterologous proteins (Fig. 3d,e,g,h). Note that the 1D ^{15}N-edited ^{1}H HSQC spectra (Fig. 3d,e) are clearly different from each other, reflecting distinct structural characteristics. The spectrum for protein XAC2775 (Fig. 3d) shows low dispersion, H_N peaks falling between 7.5 and 8.5 ppm. On the other hand, the spectrum for protein XAC0862 (Fig. 3e) shows a wide dispersion of amide proton peaks (6.5–9.5 ppm), indicative of a well-folded protein. The 2D HSQC spectra (Fig. 3g,h) of the soluble bacterial lysates confirm these observations, but in more detail. Specifically, the spectrum for protein XAC2775 contains a small number of broad lines (Fig. 3g) while that of protein XAC0862 contains a large number of sharp well-resolved lines (Fig. 3h). Additionally, the number of peaks observed in the XAC0862 2D spectrum match with that expected from the primary sequence (143). Comparison of the spectra of the XAC0862 bacterial lysate (Fig. 3b,e,h) with the purified protein (Fig. 3c,f,i) confirms that rifampicin allows the selective labelling of the heterologous protein alone as previously shown [23], and that the lysate spectra of the ^{15}N-labelled protein provides valuable information about protein conformation.

Besides XAC0862, three other *good* candidates were purified to homogeneity after labeling with ^{15}N. The HSQC spectra of these purified proteins are presented in Fig. 4. The spectra for purified XAC2000, XAC2396 and XAC3873 (Fig. 4) are very similar to those observed for their corresponding bacterial lysates. For XAC2000 and XAC2396, the observed number of peaks in the 2D spectra are in good

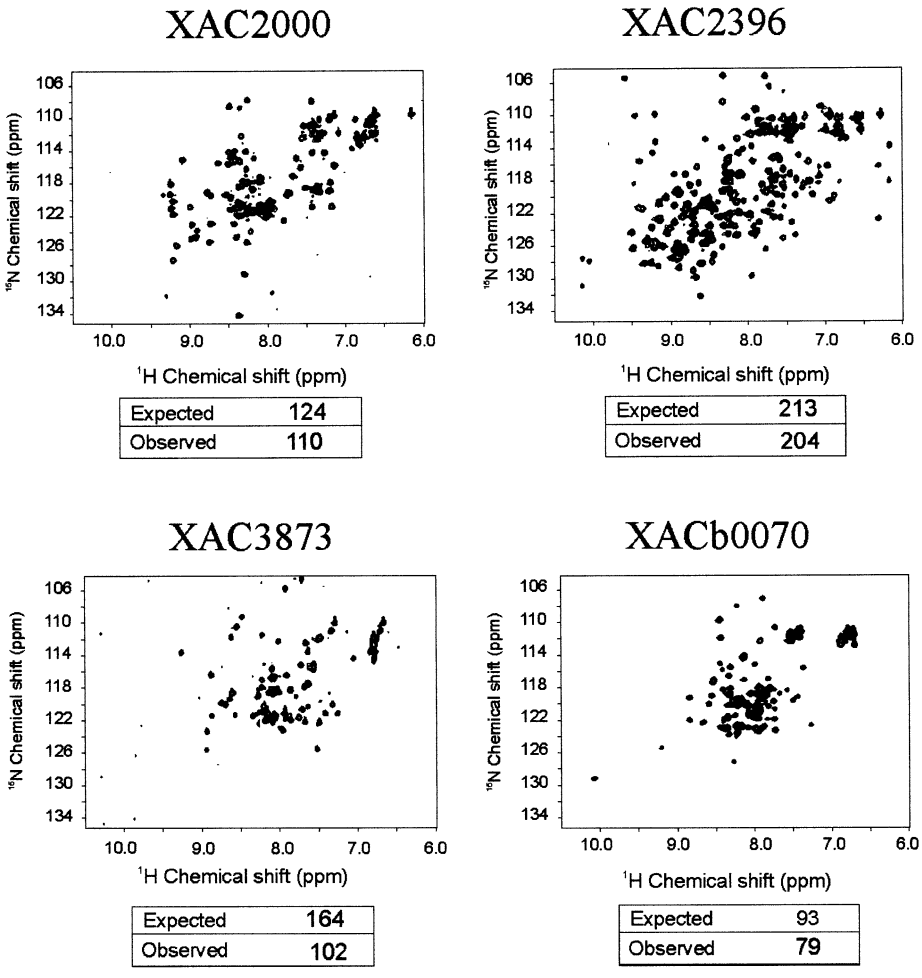

Fig. 4. 2D ^{15}N-edited ^{1}H HSQC spectra of purified proteins from ORFs XAC2000, XAC2396, XAC3873 and XACb0070. Spectra were acquired at 25°C and pH 7.0.

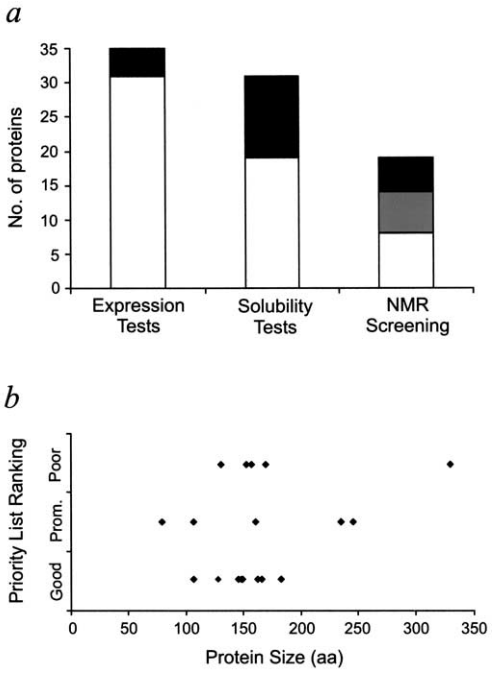

Fig. 5. a: Histogram of the number of proteins expressed (white), not expressed (black), soluble (white), insoluble (black) and with *good* (white), *promising* (gray) and *poor* (black) NMR spectra. b: Correlation between priority list ranking (good, promising and poor) and protein size (aa = number of amino acids and prom. = promising).

agreement with the expected number of peaks. These results are consistent with the hypothesis that most of the proteins presenting *good* spectra may in fact be well-folded proteins even in the bacterial lysates and that they maintain their folded state during the purification procedure. The 2D HSQC spectra of the unpurified and purified fractions of XAC3873 are also very similar and show good peak dispersion. However, in this case only 102 of the 164 expected peaks were observed in both. This may indicate overlap of resonances due to high helical content or to a well-folded domain consisting of a little more than half of the polypeptide chain. In the case of ambiguity, well-folded proteins may also be distinguished using pulse sequences that enable the identification of non-hydrogen bonded amide protons present in less structured regions of the protein and in fast exchange with water [26,36,37].

3.5. Prioritizing protein targets for further structural studies

The purpose of the screening protocol is to create a priority list to be used as a guide for choosing suitable protein candidates for the more expensive and time-consuming purification and structure determination phases of a medium- or large-scale structural biology project. The method described here allowed us to assess the conformational state of the protein in solution without the need of previous purification.

The spectra evaluated as 'promising' in Figs. 1 and 2 generally presented a lower degree of chemical shift dispersion than the *good* spectra but sharper and more intense lines than the *poor* spectra. In some cases, a narrow range of ^{1}H chemical shifts may be indicative of high α-helical content and not necessarily random coil conformations. One *promising* candidate, protein XACb0070, that showed sharp peaks but a low degree of ^{1}H chemical shift dispersion, was purified in a soluble and stable form. The 2D HSQC spectrum of the purified protein (Fig. 4) is similar to that observed in the bacterial lysate (Fig. 2b), although not the same. The number of amide peaks observed is close to that expected from its primary sequence (93). This protein aggregates at concentrations greater than 500 μM, which may contribute, in part, to the observed differences. Furthermore, there are many examples in the literature of large-scale ^{1}H and ^{15}N chemical shift changes induced by the binding of inorganic and organic cofactors [33–35]. Such factors would be present in the bacterial lysate but probably absent in the purified protein. The CD spectrum of the purified spectrum obtained was typical for an α-helical protein (data not shown).

It must be noted that during the screening procedure no protein-specific optimization whatsoever was carried out for the expression or lysis, prior to NMR analysis. Therefore proteins low in the priority list should not be discarded. Folded proteins that interact with cellular components would hamper the HSQC quality. In principle, it would be difficult to distinguish between natively unfolded proteins and misfolded proteins using our methodology [38].

Selective ^{15}N-labelling using rifampicin for NMR screening was originally described and tested using proteins whose structure and stability were previously known [23]. It was not known whether this methodology would be effective when analyzing unknown structural proteomics targets. To determine the efficiency of this method for screening targets for structural proteomics, we screened 35 previously uncharacterized *X. a.* pv *citri* proteins. In total, 42% of the tested proteins showed good NMR spectra, 32% showed promising spectra and 26% showed poor spectra (Fig. 5a). One of the concerns of using HSQC spectra to screen for well-folded proteins was that, given the size limitation of NMR, the best spectra would correspond to the smaller proteins tested whereas poor spectra would be observed for larger proteins. We nevertheless found examples of small and large proteins producing all types of spectra (*good*, *promising* and *poor*) (Fig. 5b). This shows that this method is suitable for creating a priority list for structural proteomics candidates, regardless of their size, provided they are sufficiently small for NMR analysis. 'Foldedness' defined in this way may also be of value in the screening stages of large-scale protein crystallization projects[31,32].

There are several published methods to screen promising candidates for structural proteomics projects [12,31,32,33,39]. Yee et al. [33] recently described the screening of 513 structural proteomics candidates from several genomes using efficient large-scale affinity purification of His-tagged ^{15}N-labelled proteins followed by HSQC analysis. The combination of selective ^{15}N-labelling of heterologous proteins with the rapid analysis of soluble cell lysates by NMR spectroscopy described here provides a promising alternative method for rapidly and efficiently screening structural proteomics targets for future high-resolution studies by NMR or X-ray crystallography.

Acknowledgements: We thank A.S. Pinheiro for assistance. We also thank Dr. Thelma Pertinhez for help in the acquisition of the CD and HSQC spectra of XACb0070 at the Laboratório Nacional de Luz Síncrotron in Campinas, Brazil. This work was supported with grants from FAPESP, CNPq/PRONEX (Brazil) and The Third World Academy of Science (Trieste, Italy). L.M.P.G. is supported by a PhD

studentship from FCT/MCT (Portugal). A.M.K and C.R.G. received undergraduate research fellowships from FAPESP.

References

[1] Burley, S.K. (2000) Nat. Struct. Biol. 7, 932–934.
[2] Brenner, S.E. (2001) Nat. Rev. Genet. 10, 801–809.
[3] Mittl, P.R.E. and Grütter, M.G. (2001) Curr. Opin. Chem. Biol. 5, 402–408.
[4] Stevens, R.C., Yokoyama, S. and Wilson, I.A. (2001) Science 294, 89–92.
[5] Bhattacharyya, S., Habibi-Nazhad, B., Amegbey, G., Slupsky, C.M., Yee, A., Arrowsmith, C. and Wishart, D.S. (2002) Biochemistry 41, 4760–4770.
[6] Christendat, D., Saridakis, V., Kim, Y., Kumar, P.A., Xu, X., Semesi, A., Joachimiak, A., Arrowsmith, C.H. and Edwards, A.D. (2002) Protein Sci. 11, 1409–1414.
[7] Jackson, R.M. and Russell, R.B. (2001) Comput. Chem. 26, 31–39.
[8] Service, R.F. (2002) Science 298, 948–950.
[9] Yokoyama, S. (2003) Curr. Opin. Chem. Biol. 7, 39–43.
[10] Dieckman, L., Gu, M., Stols, L., Donnelly, M.I. and Collart, F.R. (2002) Protein Exp. Purif. 25, 1–7.
[11] Boettner, M., Prinz, B., Stahl, U. and Lang, C. (2002) J. Biotechnol. 99, 51–62.
[12] Lesley, S.A. et al. (2002) Proc. Natl. Acad. Sci. USA 99, 11664–11669.
[13] Villasenor, A., Sha, M., Thana, P. and Browner, M. (2002) Biotechniques 32, 184, 186, 188–189.
[14] Lamzin, V.S. and Perrakis, A. (2000) Nat. Struct. Biol. 7 (Suppl), 978–981.
[15] Szyperski, T., Yeh, D.C., Sukumaran, D.K., Moseley, H.N. and Montelione, G.T. (2002) Proc. Natl. Acad. Sci. USA 99, 8009–8014.
[16] Bhavesh, N.S., Panchal, S.C. and Hosur, R.V. (2001) Biochemistry 40, 14727–14735.
[17] Prestegard, J.H., Valafar, H. and Tian, F. (2001) Biochemistry 40, 8677–8685.
[18] Zweckstetter, M. and Bax, A. (2001) J. Am. Chem. Soc. 123, 9490–9491.
[19] Kim, S. and Szyperski, T. (2003) J. Am. Chem. Soc. 125, 1385–1393.
[20] Guignard, L., Ozawa, K., Pursglove, S.E., Otting, G. and Dixon, N.E. (2002) FEBS Lett. 254, 159–162.
[21] Kigawa, T., Yabuki, T., Yoshida, Y., Tsutsui, M., Ito, Y., Shibata, T. and Yokoyama, S. (1999) FEBS Lett. 442, 15–19.
[22] Shimba, N., Yamada, N., Yokoyama, K. and Suzuki, E. (2002) Anal. Biochem. 301, 123–127.
[23] Almeida, F.C.L., Amorim, G.C., Moreau, V.H., Sousa, V.O., Creazola, A.T., Américo, T.A., Pais, A.P.N., Leite, A., Netto, L.E.S., Giordano, R.J. and Valente, A.P. (2001) J. Magn. Reson. 148, 142–146.
[24] Da Silva, A.C. et al. (2002) Nature 417, 459–463.
[25] Studier, F.W., Rosenberg, A.H., Dunn, J.J. and Dubendorff, J.W. (1990) Methods Enzymol. 185, 60–89.
[26] Dalvit, C. and Hommel, U. (1995) J. Magn. Reson. 109, 334–338.
[27] Nakai, K. and Horton, P. (1999) Trends Biochem. Sci. 24, 34–36.
[28] Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. (1997) Nucleic Acids Res. 25, 3389–3402.
[29] Huynen, M., Doerks, T., Eisenhaber, F., Orengo, C., Sunyaev, S., Yuan, Y.P. and Bork, P. (1998) J. Mol. Biol. 280, 323–326.
[30] Wishart, D.S. and Sykes, B.D. (1994) Methods Enzymol. 239, 363–392.
[31] Rehm, T. and Huber, R. (2002) Structure 10, 1613–1618.
[32] Montelione, G.T., Zheng, D., Huang, Y.J., Gunsalus, K.C. and Szyperki, T. (2000) Nat. Struct. Biol. 7 (Suppl), 982–985.
[33] Yee, A. et al. (2002) Proc. Natl. Acad. Sci. USA 99, 1825–1830.
[34] Zartler, E., Jenney, F.E., Terrell, M., Eidsness, M.K., Adams, M.W.W. and Prestegart, J.H. (2001) Biochemistry 40, 7279–7290.
[35] Jaren, O.R., Kranz, J.K., Sorensen, B.R., Wand, A.J. and Shea, M.A. (2002) Biochemistry 41, 14158–14166.
[36] Hwang, T.L., Mori, S., Shaka, A.J. and vanZijl, P.C.M. (1997) J. Am. Chem. Soc. 119, 6203–6204.
[37] Hwang, T.L., van Zijl, P.C.M. and Mori, S. (1998) J. Biomol. NMR 11, 221–226.
[38] Uversky, V.N. (2002) Protein Sci. 11, 739–756.
[39] Christendat, D. et al. (2000) Nat. Struct. Biol. 7, 903–909.